Inteligencja obliczeniowa - Reguły asocjacyjne

Grzegorz Madejski

2021/22

Przykład

Załóżmy, że mamy bazę danych z 10 transakcjami w sklepie. Co kupowali klienci?

T1	{masło, chleb}
T2	{chleb, ser}
T3	{masło, chleb, ser}
T4	{piwo, czipsy}
T5	{chleb, piwo}
T6	{chleb, piwo, czipsy}
T7	{masło, chleb, piwo, czipsy}
T8	{chleb, piwo, czipsy}
T9	{masło, chleb, ser, piwo}
T10	{masło, chleb, piwo, czipsy}

Przykład

Załóżmy, że mamy bazę danych z 10 transakcjami w sklepie. Co kupowali klienci?

	masło	chleb	ser	piwo	czipsy
T1	TRUE	TRUE	FALSE	FALSE	FALSE
T2	FALSE	TRUE	TRUE	FALSE	FALSE
T3	TRUE	TRUE	TRUE	FALSE	FALSE
T4	FALSE	FALSE	FALSE	TRUE	TRUE
T5	FALSE	TRUE	FALSE	TRUE	FALSE
T6	FALSE	TRUE	FALSE	TRUE	TRUE
T7	TRUE	TRUE	FALSE	TRUE	TRUE
T8	FALSE	TRUE	FALSE	TRUE	TRUE
T9	TRUE	TRUE	TRUE	TRUE	FALSE
T10	TRUE	TRUE	FALSE	TRUE	TRUE

Przykład

T1	{masło, chleb}
T2	{chleb, ser}
T3	{masło, chleb, ser}
T4	{piwo, czipsy}
T5	{chleb, piwo}
T6	{chleb, piwo, czipsy}
T7	{masło, chleb, piwo, czipsy}
T8	{chleb, piwo, czipsy}
T9	{masło, chleb, ser, piwo}
T10	{masło, chleb, piwo, czipsy}

	masło	chleb	ser	piwo	czipsy
T1	TRUE	TRUE	FALSE	FALSE	FALSE
T2	FALSE	TRUE	TRUE	FALSE	FALSE
T3	TRUE	TRUE	TRUE	FALSE	FALSE
T4	FALSE	FALSE	FALSE	TRUE	TRUE
T5	FALSE	TRUE	FALSE	TRUE	FALSE
T6	FALSE	TRUE	FALSE	TRUE	TRUE
T7	TRUE	TRUE	FALSE	TRUE	TRUE
T8	FALSE	TRUE	FALSE	TRUE	TRUE
T9	TRUE	TRUE	TRUE	TRUE	FALSE
T10	TRUE	TRUE	FALSE	TRUE	TRUE

Przykładowa reguła asocjacyjna: ”Jeśli klient kupuje masło i chleb, to kupuje też ser”. Matematycznie: {masło, chleb} $\rightarrow$ {ser} (dla drugiej tabeli: {masło=TRUE, chleb=TRUE} $\rightarrow$ {ser=TRUE})

Definicje

- Pozycje/produkty/przedmioty (ang. items) opsiują dostepne towary: $I = \{i_1, ..., i_n\}$.
- Baza transakcji T składa się z itemsetów T_i czyli zbiorów zakupionych towarów.
- Itemset to dowolny podzbiór I.
- k-itemset to podzbiór I o k elementach.
- Reguła (asocjacyjna) to para $A \rightarrow B$, gdzie A i B to itemsety.

Ocena zbiorów i reguł

- Możemy zmierzyć częstość wystepowania zbiorów (itemsetów) w bazie danych. Taką częstość nazywamy *wsparciem* (ang. support) i obliczamy wg wzoru:

$$supp(A) = \frac{\text{liczba transakcji z A jako podzbiorem}}{\text{wielkość bazy danych T}}$$

- Możemy zmierzyć prawdziwość reguł asocjacyjnych. Nazywamy to *wiarygodnością* lub ufnością (ang. confidence) i obliczamy wg wzoru:

$$conf(A \rightarrow B) = \frac{supp(A \cup B)}{supp(A)}$$

- Jeśli pytamy o wsparcie dla reguły, to chodzi o wsparcie zbioru sumującego obie strony:

$$supp(A \rightarrow B) = supp(A \cup B)$$

Zadanie

Zadanie 1

Dla podanej bazy danych podaj wsparcie i wiarygność reguł:
{chleb} → {piwo,czipsy}
{piwo} → {czipsy}
{czipsy}→ {piwo}
{masło, ser}→ {chleb}

T1	{masło, chleb}
T2	{chleb, ser}
T3	{masło, chleb, ser}
T4	{piwo, czipsy}
T5	{chleb, piwo}
T6	{chleb, piwo, czipsy}
T7	{masło, chleb, piwo, czipsy}
T8	{chleb, piwo, czipsy}
T9	{masło, chleb, ser, piwo}
T10	{masło, chleb, piwo, czipsy}

	masło	chleb	ser	piwo	czipsy
T1	TRUE	TRUE	FALSE	FALSE	FALSE
T2	FALSE	TRUE	TRUE	FALSE	FALSE
T3	TRUE	TRUE	TRUE	FALSE	FALSE
T4	FALSE	FALSE	FALSE	TRUE	TRUE
T5	FALSE	TRUE	FALSE	TRUE	FALSE
T6	FALSE	TRUE	FALSE	TRUE	TRUE
T7	TRUE	TRUE	FALSE	TRUE	TRUE
T8	FALSE	TRUE	FALSE	TRUE	TRUE
T9	TRUE	TRUE	TRUE	TRUE	FALSE
T10	TRUE	TRUE	FALSE	TRUE	TRUE

Sprzedaż

W supermarketach, po zeskanowaniu kodów kreskowych, produkty z każdego kupna są zapisane w bazie danych. Sprzedawca może sprawdzić jakie produkty są kupowane razem i dopasować do tego strategię marketingową (zestawy w promocyjnej cenie, rozmieszczenie produktów w sklepie). Angielski termin: market basket analysis.

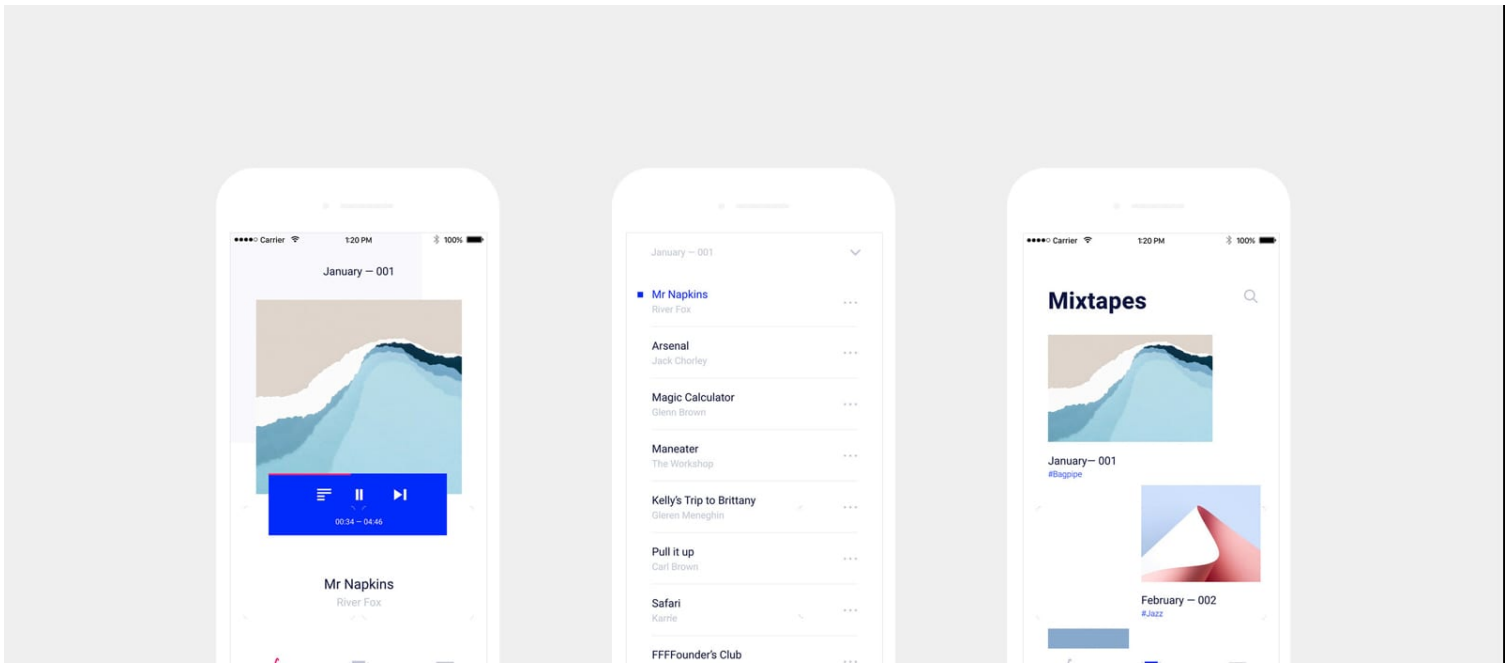
Rekomendacje

Serwisy z materiałami multimedialnymi (np.Youtube, Netflix, Spotify) mogą zbierać dane o twojej aktywności w serwisie, szukać wzorców twojego zachowania i układać reguły asocjacyjne, które będą ci rekomendowały najbardziej pasujące materiały.

Źródło obrazka: https://laptrinhx.com/how-does-netflix-know-what-movies-you-ll-enjoy-2163869420/

Projektowanie UI

Twórcy serwisów czy aplikacji mogą zbierać dane o aktywności użytkowników: gdzie klikają, do jakich menu wchodzą, gdzie scrollują. W ten sposób, mogą ulepszyć interfejs.



Źródło obrazka: https://xd.adobe.com/ideas/process/ui-design/4-golden-rules-ui-design/

Medycyna

Lekarze mogą szukać zależności pomiędzy wyśtępowaniem różnych symptomamów chorobowych, pomiędzy chorobami, symptomami a chorobami i lekami a chorobami. Takie techniki wspomagają wiedzę lekarzy na temat choroby.

Algorytm dla zbiorów częstych - pseudokod

```
function APRIORI(T, supp_min)
    Freq_1 ← {t ∈ I: supp(t) ⩾ supp_min}
    k ← 2
    while Freq_{k−1} ≠ ∅:
        Cand_k ← APRIORI-GEN(Freq_{k−1}, k)
        Freq_k ← {t ∈ Cand_k: supp(t) ⩾ supp_min}
        k ← k + 1

    return Freq_1 ∪ ··· ∪ Freq_k


function APRIORI-GEN(Freq_{k−1}, k)
    Cand_k ← ∅
    for all x, y ∈ Freq_{k−1}
        if x ≠ y and x[1 : k − 2] = y[1 : k − 2]
            z ← x ∪ y
            if (u ⊂ z for all u ∈ Freq_{k−1})
                Cand_k ← Cand_k ∪ z

    return Cand_k
```

Algorytm dla zbiorów częstych - objaśnienia

Algorytm `APRIORI`:

- wyszukuje zbiory częste (minimalny support) o wielkości jeden ($Freq_1$).
- Na podstawie $Freq_1$ generuje dobrych kandydatów ($Cand_2$) o wielkości 2
- Z kandydatów $Cand_2$ wybiera zbiory częste tworząc zbiór $Freq_2$.
- Na podstawie $Freq_2$ generuje dobrych kandydatów ($Cand_3$) o wielkości 3.
- Z kandydatów $Cand_3$ wybiera zbiory częste tworząc zbiór $Freq_3$.
- Itd. ...
- Działa tak długo, aż wyczerpie możliwości tworzenia nowych zbiorów.

Algorytm dla zbiorów częstych - przykład działania

Zakładamy, że $supp_{min} = 50\%$ (3 wystąpienia).

T_i	zakupy				
10	A	C		T	W
20		C	D		W
30	A	C		T	W
40	A	C	D		W
50	A	C	D	T	W
60		C	D	T	

	częste podzbiory
F_1	A, C, D, T, W
C_2	AC, AD,...
F_2	AC, AT, AW, CD, CT, CW, DW, TW
C_3	ACT, ACW, ATW, CDW CDT, CTW
F_3	ACT, ACW, ATW, CDW CTW
C_4	ACTW$\cdots$

Algorytm dla zbiorów częstych - objaśnienia

Funkcja `APRIORI-GEN` służy do wyłonienia dobrych kandydatów na zbiory częste o wielkości k ($Cand_k$), mając do dyspozycji zbiory częste o wielkości $k-1$ ($Freq_{k-1}$). Funkcja:

- wybiera pary x, y ze zbioru $Freq_{k-1}$ takie, żeby itemy zgadzały się na pierwszych $k-2$ miejscach. Nastepnie *łączy* takie pary. Przykład:

$$Freq_{k-1} = \{AB, AC, AD, AE, BC, BD, BE\}$$

$$Cand_k = \{ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE\}$$

- następnie *wycina* (ang. pruning) ze zbioru $Cand_k$ wszystkie elementy, których wszystkie podzbiory wielkości $k-1$ nie należą do $Freq_{k-1}$.

$$Cand_k = \{ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE\}$$

$$Cand_k = \{ABC, ABD, ABE\}$$

Algorytm do generowania reguł - objaśnienia

Mając zbiory częste możemy generować reguły. Reguły mają zadaną minimalną wiarygodność $conf_{min}$. Algorytm opiszemy słownie:

- Rozpatrujemy wszystkie możliwe zbiory częste o długości większej równej 2:
$$f \in Freq_2 \cup Freq_3 \cup ... \cup Freq_{|I|}$$
- Dla danego zbioru f rozpatrujemy wszystkie możliwe podzbiory właściwe $h \subset f$, $h \neq \emptyset$, $h \neq f$ tworząc reguły:
$$(f - h) \rightarrow h$$
- jeśli dana reguła ma wiarygodność $conf_{min}$ lub większą to jest wyświetlana, w przyciwnym wypadku odrzucana
- Powyższą procedurę można zoptymalizować: jęsli wiemy, że $AB \rightarrow CD$ jest wiarygodna, to $ABC \rightarrow D$ i $ABD \rightarrow C$ również są.

Zadanie

Zadanie 2

Dla podanej bazy danych zasymuluj działanie algorytmu Apriori do generowania zbiorów częstych. Wypisując po kolei zawartość zbiorów $Freq_1, Cand_2, Freq_2, Cand_3, Freq_3, Cand_4, Freq_4, Cand_5, Freq_5$. Przyjmij minimalne wsparcie $supp_{min} = 50\%$.

T1	{masło, chleb}
T2	{chleb, ser}
T3	{masło, chleb, ser}
T4	{piwo, czipsy}
T5	{chleb, piwo}
T6	{chleb, piwo, czipsy}
T7	{masło, chleb, piwo, czipsy}
T8	{chleb, piwo, czipsy}
T9	{masło, chleb, ser, piwo}
T10	{masło, chleb, piwo, czipsy}

	masło	chleb	ser	piwo	czipsy
T1	TRUE	TRUE	FALSE	FALSE	FALSE
T2	FALSE	TRUE	TRUE	FALSE	FALSE
T3	TRUE	TRUE	TRUE	FALSE	FALSE
T4	FALSE	FALSE	FALSE	TRUE	TRUE
T5	FALSE	TRUE	FALSE	TRUE	FALSE
T6	FALSE	TRUE	FALSE	TRUE	TRUE
T7	TRUE	TRUE	FALSE	TRUE	TRUE
T8	FALSE	TRUE	FALSE	TRUE	TRUE
T9	TRUE	TRUE	TRUE	TRUE	FALSE
T10	TRUE	TRUE	FALSE	TRUE	TRUE

Następnie podaj wszystkie reguły asocjacyjne o minimalnej wiarygodności $conf_{min} = 60\%$ i minimalnej długości 3 (itemsety wielkości 3,4,5).